International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 9 (2015) pp. 23137-23151

http://www.ripublication.com

Identifying Effective Attributes For Structuring Social Circle/Lists In Social Media Using Principal Component Analysis

[1]Samiya Shafi and [2]Harshpreet Singh

1Student-Department of computer science and engineering, Lovely Professional University, Punjab, India.
2Assisant Professor-Department of computer science and engineering, Lovely Professional University, Punjab, India.

Abstract

Big Data refers to large amount of data generated from heterogeneous resources. The data generated from social media aims in providing information about the individual and their respective behavior. The user attributes in the social data also provides a structure to form various social networks. This motivation of the work carried out in this paper focuses on analyzing the social data for finding attributes which are effective in making circles and lists in social media. Using various dimensions such as name, school, place, location, birthday, degree, class, school, name, hometown etc the paper aims in providing the principle components which aids in building the social network. The paper employs Principal Component Analysis for the analysis of social media data for Circle/List formation. Principal Component Analysis is mathematical procedure to reduce number of dimensions of dataset by maintaining its original variability. The approach is utilized for isolating the attribute effective in making of circles. Interpretation and validation of the proposed methodology will be plotted by scree test.

Keywords: Variability, dimension, Principal Component Analysis, scree test

Introduction

Big Data refers to large, complex and growing volume of data sets with evolving relationships from heterogeneous and autonomous sources[1].Big Data is generated from number of data intensive application like online discussions, Flicker (public picture sharing site), sensors, online shopping sites, social media giants (Facebook, Twitter, LinkedIn, YouTube, Google+ and more),scientific data analysis, mobile

devices and more[2]. Each day Google has 1 billion above queries, Twitter has 250 million above tweets, Facebook has 800 million above updates, and YouTube has more than 4 billion views per day [3]. Over 2.5 quintillion bytes of data are generated everyday and 90% of the present data has been created in the last two years [4].

Social networking sites have hundreds of millions of users. Social networks are tool for connecting people, and mirror real-life relationships and society. Users of social media maintain profile information like location, education; concentration, birthday, education; class; school, first_name, hometown, and much more [5]. Cost and overhead previously make communication unfeasible, but advances in social networking technology have made sharing possible[6]. Social media enable users to communicate information to anybody who is looking for it. Social media sites like Google+, Facebook, Twitter and so on, have large number of users. Google+ has 540 million monthly active users[7] and sharing 1.5 billion photos each week. Google+ is fifth social networking site in world. Facebook currently has over 650 million active users every day. Twitter generates half a billion tweets-per day. Social Network is powerful means of sharing, organizing and finding contents, contacts[8].Users in social networking sites join network, create profiles and relationships with any users of same social network with whom they associate. Social networks revolve around users, user group friends into circle in Google+ and list in Facebook.

Social networking sites allow users to categorize their friends manually into their social groups, circles and in their social list. Manually categorization of friends in lists and circles is time consuming and lengthy [9].Circles can be used for content filtering, privacy and sharing data between its users[10]. Data is collected from social networking sites about user through different attributes like "tag", "comments", "like", "status", "photos" and "video" which is refer as social media data. These data are the basis for creating models of the relationships between users. They can be used to significantly increase the relevance of what is shown to the user, for advertising and marketing purposes of products [11].

This paper focus on user attributes used as profile information which help user to form social circle. User add friends into circles and list by categorizing them using profile information and making it easy to filter the friends. This paper presents an approach on finding attributes which effectively contribute in making the circle using Principal Component Analysis.

Let us consider *n* as numbers of attributes which define a user in social media for creating circles. Principal Component Analysis (PCA) helps to identify attributes which are meaningful and reduce dimensionality of data. With PCA the complexity of data can be reduced, need less number of plots to analyze. Principal Component Analysis is mathematical procedure to reduce number of dimensions (attributes) of dataset but maintaining its original variability. Interpretation and validation of the proposed methodology will be plotted by scree test[12].

Instead of working on all m-dimensions, we will first perform PCA on original data (m-dimensions) then use only first few components say PC1, PC2 in analysis. These reduced dimensions can be used to focus on making circle. Circle is made with users, users have number of attributes focusing on all the attributes is very lengthy,

costly[13]. With PCA reduced attributes, analysis can be easy, and focusing on only those attributes which are efficient in making circles.

This paper is organized as follows. Section 2 gives an overview of the mathematical foundations of Principal Component Analysis. Section 3 briefs about the dataset which is used to conduct Principal Component analysis. Section 4 describes the result of implementation of Principal Component Analysis on social media data. The PCA reduce dimensions of original data. And describes which attribute is effective in making circles. Section 5 provides the conclusion of paper and future research directions.

Principal Component Analysis

Principal Component Analysis is a procedure that transforms number of correlated variables into smaller number of uncorrelated variables called principal components[14].The goal is to reduce number of variables but retain original meaning of the data[15],[16]. PCA is equivalent in finding direction axis which have maximum variance, then using these new directions to define new basis.

Let us consider a multivariate dataset that is represented in terms of a $n \times m$ matrix, $A_{n,m}$ where m columns are sample and n rows are variables.

$$A_{n,m} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,m} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,m} \end{bmatrix}$$

Then matrix A is linearly transform into another matrix B of same dimension $n \times m$,so that for some $n \times n$ matrix, C.

$$B = CA \qquad \text{Eq. 1}$$

The above equation represents change of basis. Let us consider the row of C to be row vectors c_1, c_2, $c_3 \ldots c_n$ and column of A to be column vectors a_1, a_2... a_m then Eq. 1 can be represented as

$$B = CA = \left(Ca_1, Ca_2, Ca_3, \ldots Ca_n\right) = \begin{bmatrix} c_1a_1 & c_1a_2 & \cdots & c_1a_m \\ c_2a_1 & c_2a_2 & \cdots & c_2a_m \\ \vdots & \vdots & \ddots & \vdots \\ c_na_1 & c_na_2 & \cdots & c_na_m \end{bmatrix}$$

The $c_i a_j \subset R^n$ and $c_i a_j$ is Euclidean dot product. The $c_i a_j$ representation means that original data, A is projected on to rows of C. Now, the columns of *C* that is (c_1, c_2, $c_3 \ldots c_n$) are new basis for showing rows of *A*. The columns of C then are principal component directions.

In order to represent the independence between principal components in new basis the variance of the data in the original basis is considered. Original data is de-correlate by finding the direction in which variance is maximized .These direction are

used to define the new basis as defined by new basis *C*. The variance of random variable, *W* with mean μ is given by following equation

$$\sigma_W^2 = \mathrm{E}\left(W - \mu \right)^2 \quad \text{Eq. 2}$$

Where W=random variable, μ=mean, σ_W = variance of W random variable. Considering a vector $\tilde{u}$ of n discrete measurements, $\tilde{u} = \tilde{u}_1, \tilde{u}_2, \tilde{u}_3 ..... \tilde{u}_n$ with mean of u vector is μ_u .A translated set of measurements $u = u_1, u_2, u_3 ..... u_n$ can be obtained by subtracted mean from each measurement that has zero mean. The variance of these the measurement u can be given by following equation

$$\sigma_u^2 = \frac{1}{n-1} \left(u^T \right) \quad \text{Eq. 3}$$

If second vector $v = (v_1, v_2, v_3 ... v_n)$ of n measurements again with zero mean is considered then it can be generalized to obtain covariance by considering both u and v as given in Eq. 4. *A c*ovariance can be used to measure how much two variables change together, covariance can be positive or negative. In PCA covariance is positive covariance, close to zero. Variance is special case of covariance, when variables are similar. In PCA variance are maximized and covariance are minimized. The variance between u and v vector is as follows:

$$\sigma_{uv}^2 = \frac{1}{n-1} \left(v^T \right) \quad \text{Eq. 4}$$

Generalize the matrix $A_{n,m}$ can be used to represent data matrix.

$$A_{n,m} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,m} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,m} \end{bmatrix} = [\boldsymbol{a_1} \quad \boldsymbol{a_2} \quad \cdots \quad \boldsymbol{a_m}] \in R^{n \times m}, a_i^T \in R^m \quad \text{Eq. 5}$$

Column vector for each variable contain samples for one particular variable. Here a_i is vector for i[th] variable of *n* samples. Covariance matrix,$m \times m$ can be computed using following equation

$$C_A = \frac{1}{n-1} AA^T = \frac{1}{n-1} \begin{bmatrix} a_1 a_1^T & a_1 a_2^T & \cdots & a_1 a_m^T \\ a_2 a_1^T & a_2 a_2^T & \cdots & a_2 a_m^T \\ \vdots & \vdots & \ddots & \vdots \\ a_m a_1^T & a_m a_2^T & \cdots & a_m a_m^T \end{bmatrix} \in R^{m \times m} \quad \text{Eq. 6}$$

where C_A is covariance.

The covariance between variable *m* is computed using the above equation. Covariance matrix is symmetric and square matrix. Compute all covariance pairs between m variables. Off-diagonal values are covariance's between m pairs and principal diagonal values are variances.

The dataset can be linearly transformed , A into B using equation *B=CA*. Make some supposition feature that transformed matrix, *B* to exhibit and relate supposition features to covariance matrix C_B.Supposition is that transformed matrix, *B* have

uncorrelated variables ,that is matrix C_B have covariance of variables as possible close to zero. Requirement for covariance matrix C_B are to maximize variance and to minimize off-diagonal covariance between variables. Result from this supposition is that find covariance close to zero, C_B .Choose transformation matrix, C with principal diagonal C_B ,then objectives of PCA are achieved.

Consider new supposition that vector $c_1, c_2, c_3, \ldots . c_n$ are orthogonal. The covariance matrix C_B and B can be represented as follows

$$C_B = \frac{1}{n-1} BB^T = \frac{1}{n-1} (CA)(CA)^T = \frac{1}{n-1} (CA)(C^T A^T) = \frac{1}{n-1} C(AA^T)C^T$$

$$C_B = \frac{1}{n-1} CXC^T \qquad \text{Eq. 7}$$

where $X = AA^T$, X is $m \times m$ symmetric matrix, $(AA^T)^T = (A^T)^T (A)^T = AA^T$ Theorem of linear algebra is applied which states that every square symmetric matrix is orthogonally diagonalisable. Theorem can be represented as

$$X = GPG^T \qquad \text{Eq. 8}$$

Where G is $m \times m$ orthogonal matrix whose columns are orthogonal eigenvectors of X, and principal diagonal, P entries are eigenvalues of X. The highest eigen value is first principal component (PC1), the second highest is second principal component (PC2), and so on[17].

After computing eigenvalues and eigenvectors of $X = AA^T$, sort eigenvalues in descending order and place these eigenvalues on diagonal entries, P. Now construct an orthogonal matrix, G by placing eigen vectors with highest eigen value in first column, the eigenvector of second highest in second column and so on[18],[19].The objective of diagonalising covariance matrix, of transformed data is achieved. The principal components (columns of C) are eigen vectors of covariance matrix, AA^T and columns of P are in decreasing order of 'importance'. The first column is more meaning full as compared to other columns.

Implementation Details

The objective of the paper is to apply the Principal Component Analysis (PCA) method on social media dataset to find attributes which help in making particular circles.

This paper considers the dataset with 244 attributes which are used for building the profile of the user in Google+[20]. The dataset consists of number of circles, their users and attributes of the users. Attribute value are either '1' or '0', '1' means attribute consists of a value,'0' means attribute does not contain any value. The attribute values are anonymized for privacy concerns. Profile information of a user is represented in tree structured as given in figure 1 which consists of the attributes as described in the dataset.

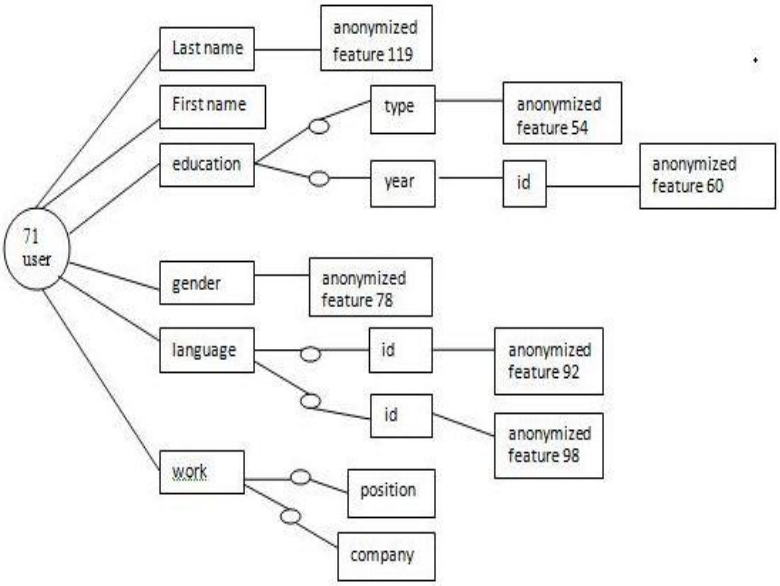

Figure 1: User71 attribute tree .The attribute values are presented as in table 1, where the value of the attributes are marked '1' or '0' accordingly.

Table 1: Attribute Present in user71 and attribute name

Attribute Present	**Attribute Name**
1	Lastname:anonymizedfeature119
0	First name
1	Education:type:anonymizedfeature54
1	Education:year:id:anonymizedfeature60
1	Gender:anonymizedfeature78
1	Language:id:anonymizedfeature92
1	Language:id:anonymizedfeature98
0	Work: position
0	Work: company

The dataset consists of 24 circles. Every circle is formed from at least 1 user and ranging till 133 users as given in table 2. Let us take an example from dataset **circle1** contains only **one user** and **circle15** contains **133 users**. The same user can be in different circles or in one circle. **User258** in dataset is present in **circle4** and **circle16**.Circles tend to overlap is they comprise of users with similar id. As in the dataset circle 1 and 16, 8 and 20 overlap each other as shown in table 3.

Table 2: Circle names with their user count

Circle name	**Circle 0**	**Circle 1**	**Circle 2**	**Circle 3**	**Circle 4**	**Circle 5**	**Circle 6**	**Circle 7**	**Circle 8**	**Circle 9**	**Circle 10**	**Circle 11**
No. of users in circle	20	1	9	3	17	1	20	2	1	10	4	30
Circle name	**Circle 12**	**Circle 13**	**Circle 14**	**Circle 15**	**Circle 16**	**Circle 17**	**Circle 18**	**Circle 19**	**Circle 20**	**Circle 21**	**Circle 22**	**Circle 23**
No. of users in circle	1	5	2	133	32	9	1	13	6	1	1	3

Table 3: Circle with their users

Circle name	**User ids of circles**
Circle 4	125,344,295,257,55,122,223,59,268,280,84,156,***258***,236,250, 239
Circle 16	251,94,330,5,34,299,254,24,180,194,281,101,266,135,197,***17 3***,36,9,85,57,37,***258***,309,80,139, 202,187,249,58,127,48,92
Circle 1	***173***
Circle 8	***282***
Circle 20	244***,282***,262,293,220,174

The **circle8** present in dataset contains only one user that is **user282** and another **circle20** has many users, one user among **circle20** is **user282** as shown in table 3. The users which are not connected in form of friends or having a common attribute cannot form the circle. The user should be connected to each other directly or indirectly. The study shows PCA reduce the dimensionality of social data. The dataset consists of 224 attributes for each user, making it difficult to work on all 224 attributes in formation of circle. In order to analyze the data PCA provides a way to find meaningful attributes which have effective contribution in making circles.

Considering the above dataset, Let us take circle0 for our analysis. Circle0 consists of 20 users with the user ids as user71, user215, user54, user61, user298, user229, user81, user253, user193, user97, user264, user29, user132, user110, user163, user259, user183, user334, user245, user222.

$$\operatorname{cov}\left(user, attributes\right) = \begin{array}{c} \\ 0attr \\ 1attr \\ 2attr \\ 3attr \\ 4attr \\ 5attr \\ 6attr \\ 7attr \\ \vdots \\ 216attr \\ 217attr \\ 218attr \\ 219attr \\ 220attr \\ 221attr \\ 222attr \\ 223attr \end{array} \begin{array}{c} \begin{array}{cccccccccccc} 0attr & 1attr & 2attr & 3attr & 4attr & 5attr & 6attr & 7attr & \cdots & \cdots & 223attr \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.9 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.005 & \cdots & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.04 & \cdots & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \vdots & \vdots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \vdots & \vdots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -0.005 & \vdots & \vdots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \vdots & \vdots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \vdots & \vdots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \vdots & \vdots & 0 \end{bmatrix} \end{array}$$

The covariance matrix for circle0 can be computed using Equation 6. The total number of attributes of user profile is 224. So, the matrix for circle0 will be 224x224 matrix. After finding covariance matrix, the eigenvalues are computed using Equation8. The eigen value of all attribute and their variance is as given in Table4.

Table 4: Variance of component in principle component

Attr_no	Attribute/ Dimension	Eigen-values	Variance %	Cumul-ative %
Attr92	92 languages;id;anonymized feature 92	0.247	3.8506	3.85
Attr53	53education;type;anonymized feature 53	0.247	3.8506	7.7006
Attr55	55education;type;anonymized feature 55	0.2475	3.8506	11.5512
Attr77	77gender;anonymized feature 77	0.2475	3.8506	15.4018
Attr78	78gender;anonymized feature 78	0.2475	3.8506	19.2524
Attr65	65 education; year; id;anonymized feature 65	0.21	3.267	22.5194
Attr50	50 education; school; id;anonymized feature 50	0.21	3.267	25.7864
Attr54	54education;type;anonymized feature 54	0.1875	2.917	28.7034
Attr52	52 education; school; id;anonymized feature 52	0.1275	1.983	30.6864
Attr59	59 education;year;id;anonymized feature59	0.1275	1.983	32.6694
Attr63	63education;year;id;anonymized feature63	0.1275	1.983	34.6524
Attr126	126locale;anonymized feature 126	0.1275	1.983	36.6354
Attr127	127locale;anonymized feature 127	0.1275	1.983	38.6184
Attr128	128location;id;anonymized feature 128	0.1275	1.983	40.6014
Attr1138	138location;id;anonymized feature 137	0.1275	1.983	42.5844
Attr 141	141work;employer;id;anonymized feature 140	0.1275	1.983	44.5674
Attr7	7birthday;anonymized feature 7	0.09	1.4	45.9674
Attr14	14education;concentration;id;anonymized feature 14	0.09	1.4	47.3674
Attr90	90languages;id;anonymized feature 90	0.09	1.4	48.7674
Attr98	98languages;id;anonymized feature 98	0.09	1.4	50.1674
Attr100	100languages;id;anonymized feature 100	0.09	1.4	51.5674
Attr129	129location;id;anonymized feature 129	0.09	1.4	52.9674
Attr133	133location;id;anonymized feature 133	0.09	1.4	54.3674
Attr160	160work;end_date;anonymized feature 157	0.09	1.4	55.7674
Attr156	156work;employer;id;anonymized feature 52	0.09	1.4	57.1674
Attr165	165work;end_date;anonymized feature 162	0.09	1.4	58.5674

Attr169	169work;end_date;anonymized feature 166	0.09	1.4	59.9674
Attr171	171work;end_date;anonymized feature 168	0.09	1.4	61.3674
Attr173	173work;end_date;anonymized feature 170	0.09	1.4	62.7674
Attr181	181work;location;id;anonymized feature 176	0.09	1.4	64.1674
Attr182	182work;location;id;anonymized feature 177	0.09	1.4	65.5674
Attr200	200work;position;id;anonymized feature 193	0.09	1.4	66.9674
Attr206	206work;start_date;anonymized feature 160	0.09	1.4	68.3674
Attr215	215work;start_date;anonymized feature 201	0.09	1.4	69.7674
Attr217	217work;start_date;anonymized feature 202	0.09	1.4	71.1674
Attr32	23education;degree;id;anonymized feature 23	0.0475	0.739	71.9064
Attr32	32education;school;id;anonymized feature 32	0.0475	0.739	72.6454
Attr43	43education;school;id;anonymized feature 43	0.0475	0.739	73.3844
Attr58	58education;year;id;anonymized feature 58	0.0475	0.739	74.1234
Attr60	60education;year;id;anonymized feature 60	0.0475	0.739	74.8624
Attr66	66education;year;id;anonymized feature 66	0.047	0.739	75.6014
Attr68	68education;year; id;anonymized feature 68	0.047	0.739	76.3404
Attr84	84hometown;id;anonymized feature 84	0.047	0.739	77.0794
Attr94	94languages;id;anonymized feature 94	0.0475	0.739	77.8184
Attr103	103languages;id; anonymized feature 103	0.0475	0.739	78.5574
Attr 106	106last_name;anonymized feature 106	0.0475	0.739	79.2964
Attr118	118last_name;anonymized feature 118	0.0475	0.739	80.0354
Attr134	134 location;id;anonymized feature 134	0.047	0.739	80.774
Attr139	139 location;id;anonymized feature 138	0.0475	0.739	81.5134
Attr144	144work;employer; id;anonymized feature 143	0.0475	0.739	82.2524
Attr148	148work;employer; id;anonymized feature 147	0.0475	0.739	82.9914
Attr149	149work;employer; id;anonymized feature 50	0.0475	0.739	83.7304
Attr150	152 work;employer; id;anonymized feature 150	0.0475	0.739	84.4694
Attr153	153work;employer; id;anonymized feature 151	0.0475	0.739	85.2084
Attr164	164work;end_date; anonymized feature 161	0.0475	0.739	85.9474
Attr172	172work;end_date; anonymized feature 169	0.0475	0.739	86.6864
Attr174	174work;end_date; anonymized feature 171	0.0475	0.739	87.4254
Attr66	66education;year; id;anonymized feature 66	0.0475	0.739	88.1644
Attr68	68education;year; id;anonymized feature 68	0.0475	0.739	88.9034
Attr84	84hometown;id; anonymized feature 84	0.0475	0.739	89.6424
Attr175	175work;end_date; anonymized feature 172	0.0475	0.739	90.3814
Attr179	179work;location; id;anonymized feature 132	0.0475	0.739	91.1204
Attr191	191 work;position; id;anonymized feature 184	0.0475	0.739	91.8594
Attr192	192 work;position; id;anonymized feature 185	0.0475	0.739	92.5984
Attr195	195 work;position; id;anonymized feature 188	0.0475	0.739	93.3374
Attr201	201work;start_date;anonymized feature 157	0.0475	0.739	94.0764
Attr202	202work;start_date;anonymized feature 194	0.0475	0.739	94.8154
Attr207	207 ork;start_date ;anonymized feature 197	0.0475	0.739	95.5544
Attr210	210work;start_date ;anonymized feature 164	0.0475	0.739	96.2934
Attr211	211work;start_date;anonymized feature 199	0.0475	0.739	97.0324
Attr213	213work;start_date;anonymized feature 166	0.0475	0.739	97.7714
Attr214	214work;start_date;anonymized feature 200	0.0475	0.739	98.5104
Attr216	216work;start_date;anonymized feature 168	0.0475	0.739	99.2494
Attr220	220work;start_date;anonymized feature 171	0.0475	0.739	99.9884

The attributes '92 languages; id; anonymized feature 92','53 education; type; anonymized feature 53','55education; type; anonymized feature 55 ', '77 gender;

anonymized feature 77', '78gender;anonymized feature 78','65 education; year; id; anonymized feature 65', '50 education; school; id; anonymized feature 50','54education;type;anonymized feature 54','52 education; school; id; anonymized feature 52'etc have major contribution in making of the circle 0, and the attributes which have zero contribution in making this circle are '0 birthday; anonymized feature 0', '1 birthday; anonymized feature 1',' 2 birthday; anonymized feature 2',' 48 education; school; id; anonymized feature 48',' 70 education; year; id; anonymized feature 70',' 87 hometown; id; anonymized feature 87',' 86 hometown; id; anonymized feature 86', etc.

The study shows that first eighteen attributes explained 45% of total variability of data attributes. Portion of first four different eigen values are 19%, 7%, 3%, 16% as shown in

Figure 2.

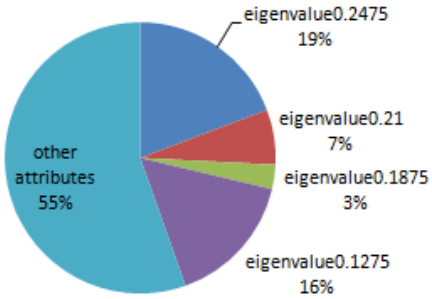

Figure 2: Total Variability of Data Attributes

The scree test plots the eigenvalues with respect to their component, the large eigenvalues and small eigenvalues display "break" between components. The attributes which appear before the "break" are supposed to be meaningful and those attributes which appear after the break are supposed to be unimportant and are not retained. If the scree test shows number of large breaks, in that case attributes appearing before last large break are supposed to be important.

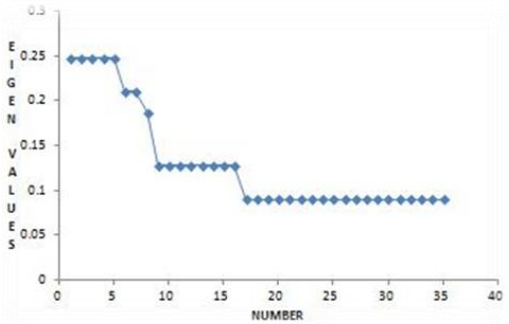

Figure 3: Scree plot of Principal Component

Scree plot of principal component analysis of 74 variables (out of 224 variables, only 74 have eigenvalues, rest have zero value) are shown in
Figure 3, presents number of big breaks before eigenvalues begin to level down. The Scree test shows that first nine components are significant for the formation of circle. Structure of meaningful components are given in table 5.

Table 5: Structure of First Nine Components Dataset

Attribute	Component1	Component2	Component3	Component4	Component5	Component6	Component7	Component8	Component9
53attr	1	0	0	0	0	0	0	0	0
55attr	0	1	0	0	0	0	0	0	0
77attr	0	0	1	0	0	0	0	0	0
78attr	0	0	0	1	0	0	0	0	0
92attr	0	0	0	0	1	0	0	0	0
50attr	0	0	0	0	0	1	0	0	0
65attr	0	0	0	0	0	0	1	0	0
52attr	0	0	0	0	0	0	0	1	0
54attr	0	0	0	0	0	0	0	0	1

Each Meaningful Component is explained as below:

Component 1 is eigenvector of first eigen value. The main part of first component is **53attr**. Thus, this component can provide a great grouping among users (user61, user229, user81, user253, user264, user264, user29, user110, user183, user334, user245, user222) from the aspect of **53attr** that is '**53 education; type; anonymized feature 53**' .

Component 2 is eigenvector of 2nd eigen value. The second component is affected by attribute '**55 education; type; anonymized feature 55**'.The second component provide grouping among users(user61, user229, user81, user253, user264, user29,user110, user183, user334, user245, user222) form aspect of **55attr**.

Component 3 is eigen vector of third eigen value. The **attr77** is main part of component 3.Third component provides grouping among users(user61, user229, user81, user253, user264, user29, user110, user183, user334, user245, user222) from aspect of attr77 '**77 gender; anonymized feature 77**'.

Component 4 is eigen vector of fourth eigen value. The **attr78** is main part of component 4.Fourth component provides grouping among users(user71, user215, user54, user297, user193, user97, user132, user163, user259) from aspect of **'78 gender; anonymized feature 78'**.

Component 5 is eigen vector of fifth eigen value. The **attr92** is main part of component five. Fifth component provides grouping among users (user71, user229, user81, user193, user132, user183, user334, user222) from aspect of **' 92 languages; id; anonymized feature 92'**.

Component 6 is eigen vector of sixth eigen value. The **attr50** is main part of component six. Sixth component provides grouping among users (user297, user229, user29, user334, user222) from aspect of **' 50 education; school; id; anonymized feature 50'**.

Component 7 is eigen vector of seventh eigen value. The **attr65** is main part of component seven. Seventh component provides grouping among users (user71, user297, user229, user183, user334, user222) from aspect of attribute **' 65 education; year; id; anonymized feature 65'**.

Component 8 is eigen vector of eighth eigen value. The **attr52** is main part of this component. This component provides grouping among users (user81, user163, user183) from aspect of attribute ' **52 education; school; id; anonymized feature 52**'.

Component 9 is ninth eigen vector of ninth eigen value. The **attr54** is main part of this component. This component provides grouping among users (user81, user163, user71, user253, user215, user245) from aspect of attribute ' **54 education; type; anonymized feature 54**'.

The **attr53, attr55, attr77, attr78, attr92** groups maximum number of users and they also provide overlapping of users in groups as shown in Table 6.

Table 6: Representation of major attributes of components with their users,'1' represents user is in group and '0' represents particular user is not in group

Users	53atttr	55attr	77attr	78attr	92attr	50attr	65attr	52attr	54attr
71user	1	0	0	1	1	0	1	0	1
215 user	0	0	0	1	0	0	0	0	1
54 user	0	0	0	1	1	0	0	0	0
61 user	1	0	1	0	0	0	0	0	0
297 user	1	1	0	1	0	1	1	0	0
229 user	1	1	1	0	1	1	1	0	0
81 user	1	1	1	0	1	0	0	1	1
253 user	1	0	1	0	0	0	0	0	0
193 user	0	1	0	1	1	0	0	0	0
97 user	0	0	0	1	0	0	0	0	0
264 user	0	0	1	0	0	0	0	0	0
29 user	1	1	1	0	0	1	0	0	0
132 user	1	1	0	1	1	1	0	0	0
110 user	0	0	1	0	0	0	0	0	0
163 user	0	0	0	1	0	0	0	1	1
259 user	0	0	0	1	0	0	0	0	0
183 user	1	1	1	0	1	0	1	1	0
334 user	1	1	1	0	1	1	1	0	0
245 user	0	0	1	0	0	0	0	0	1
222 user	1	1	1	0	1	1	1	0	0

Before applying PCA all 224 attributes are to be considered and which attribute is meaningful in forming the circle is difficult to judge. But after the implementation of PCA only few attributes can be considered in forming the circles.

At the starting of paper, we have 224 attributes. After the implementation of Principal Component Analysis on data, the number of attributes are reduced to only 5 attributes. These 5 attributes are '**53 education; type; anonymized feature 53**', '**55 education; type; anonymized feature 55**', '**77 gender; anonymized feature 77**', '**78 gender; anonymized feature 78**', ' **92 languages;id; anonymized feature 92**'.These attributes are efficient in making a circle.

Firstly, PC1 and PC2 are plotted in

Figure 4, PC1 is x-axis and PC2 is y-axis. Plotting shows that maximum numbers of users are plotted in x-axis direction. Secondly, PC2 and PC1 are plotted in

Figure 5, PC2 is x-axis and PC1 is y-axis. Plotting shows that minimum numbers of users are in x-axis direction. The conclusion is that PC1 is axis in which majority of users vary as compared to PC2.

Conclusion

Large numbers of data are produced by social networking sites and number of users create and access their accounts daily. The Social networking sites analyze interest of online user and show advertisements on user account according to their interested. The social networking site analyses all attributes of users to find similarity and interest. PCA can reduced job of working on all attributes and provides only important attribute which are effective.

Principal component analysis (PCA) is very powerful mathematical tool for reducing number of dimensions that are very effective in making circle. PCA can reduce effort and cost by reducing unimportant variables. Our future research will focus on analyzing PCA reduced attributes of circles with clustering techniques to understand the effects of PCA in reducing the effort of clustering algorithms.

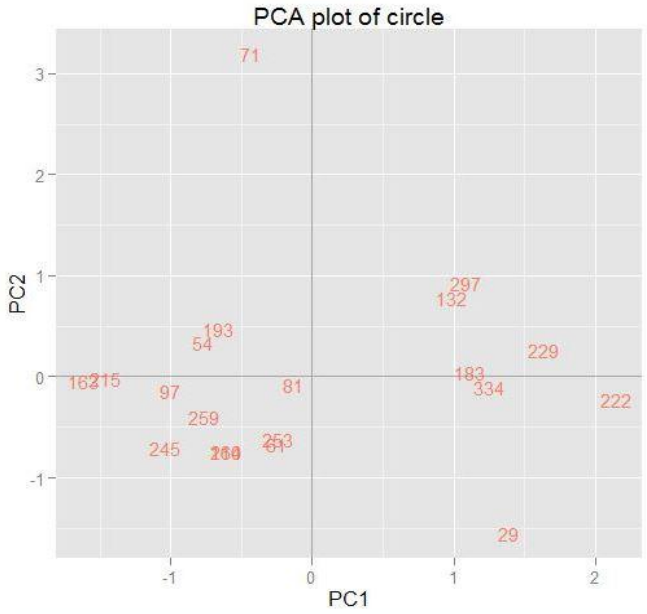

Figure 4: PC1 on x-axis and PC2 on y-axis

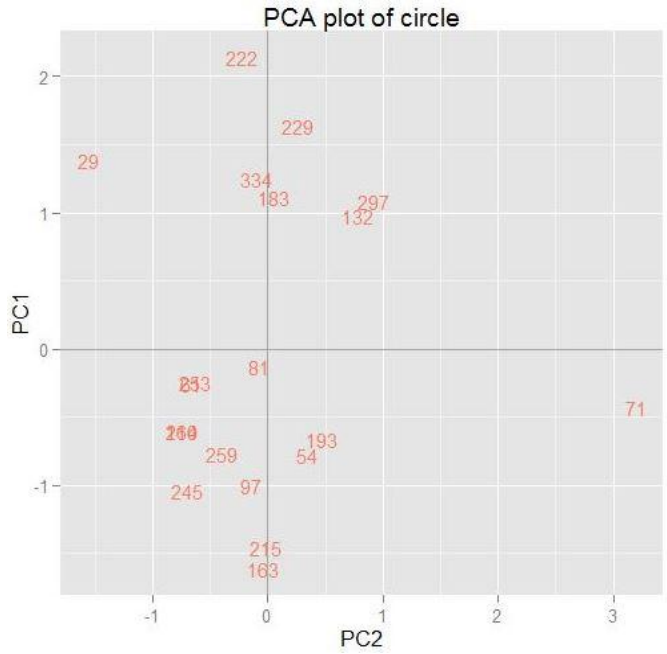

Figure 5: PC2 on x-axis and PC1 on y-axis

References

[1] Wu.X, Zhu.X, Wu.G.Q, Ding.W, "Data mining with big data, *Knowledge and Data Engineering" IEEE Transactions on 26*(1), 97-107, 2014.

[2] J. K. Laurila,D.Gatica-Perez, "The mobile data challenge: Big data for mobile computing research", In Pervasive Computing (No. EPFL-CONF-192489),2012.

[3] B. Thakur,M. Mann," Data Mining for Big Data", International Journal of Advanced Research in Computer Science and Software Engineering Volume 4 Issue 5:469-473, 2014.

[4] L. Wang, J.Zhan ,"A big data benchmark suite from internet services", arXiv preprint arXiv:1401.1406 ,2014.

[5] N. B. Ellison," Social network sites: Definition, history, and scholarship",*Journal of Computer-Mediated Communication*. *13*(1), 210-230, 2007.

[6] Bae, Jonghoon,M.Insead M,"Partner substitutability, alliance network structure, and firm profitability in the telecommunications industry",*Academy of Management Journal* 47.6: 843-859,2004.

[7] M. Meeker," Internet trends 2014-code conference"*Retrieved May*, *28*, 2014.

[8] A. Mislove," Measurement and analysis of online social networks", 2007.

[9] J. Mcauley , J. Leskovec J," Learning to discover social circles in ego networks",2012.

[10] G. Smith ,R. Boreli , M. A. Kaafar," A Layered Secret Sharing Scheme for Automated Profile Sharing in OSN Groups", In *Mobile and Ubiquitous Systems: Computing, Networking, and Services* .pp. 487-499. Springer International Publishing. 2014.

[11] R. Hochreiter , C. Waldhauser, "Data Mining Cultural Aspects of Social Media Marketing. In *Advances in Data Mining. Applications and Theoretical Aspects"*.pp. 130-143. Springer International Publishing, 2014.

[12] A B Costello," Getting the most from your analysis. *Pan", 12*(2), 131-146,2009.

[13] S. Brauer, T. C. Schmidt," Are Circles Communities? A Comparative Analysis of Selective Sharing in Google+", In *Distributed Computing Systems Workshops (ICDCSW), 2014 IEEE 34th International Conference on* June . pp. 8-15. IEEE, 2014.

[14] B. Moore," Principal component analysis in linear systems: Controllability, observability, and model reduction", *Automatic Control, IEEE Transactions on 26*(1), 17-32, 1981.

[15] J.Shlens ,"A tutorial on principal component analysis", *arXiv preprint arXiv:1404.1100*, 2014.

[16] S. Kolenikov ,G. Angeles," The use of discrete data in principal component analysis for socio-economic status evaluation", 2005,*Retrieved Nov*, *21*, 2013.

[17] F. B. Lukibisi ,T. Lanyasunya," Using principal component analysis to analyze mineral composition data", In *Proceedings of the 12th kari (kenya agricultural research institue) biennial scientific conference"*, November.pp. 8-12 ,2010.

[18] P. Comon , "Independent component analysis, a new concept?",*Signal processing, 36*(3), 287-314,1994.

[19] H Abdi , L. J. Williams," Principal component analysis", *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(4), 433-459 ,2010.

[20] J. Leskovec , A. Krevl, "SNAP Datasets: Stanford large network dataset collection", 2014.